Phylogenetic Analysis

Caro-Beth Stewart, Ph.D.
Associate Professor
Department of Biological Sciences
University at Albany, SUNY
Albany, New York 12222
c.stewart@albany.edu

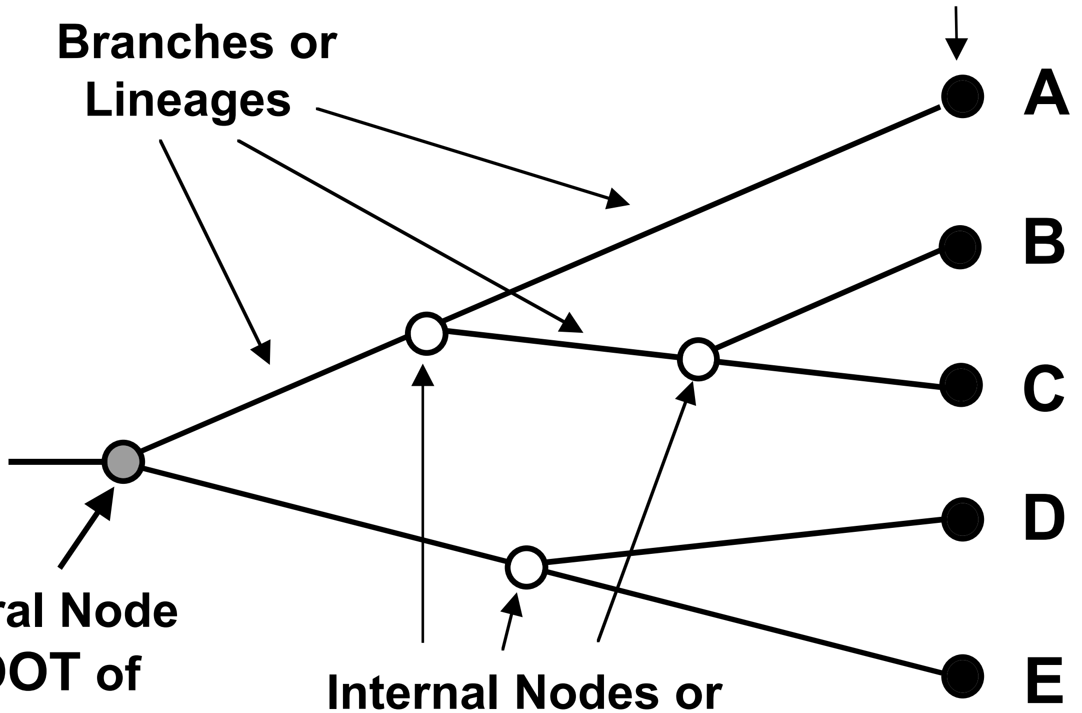
What *is* phylogenetic analysis and *why* should we perform it?

Phylogenetic analysis has two major components:

1. **Phylogeny inference or “tree building”** — the inference of the branching orders, and ultimately the evolutionary relationships, between “taxa” (entities such as genes, populations, species, etc.)

2. **Character and rate analysis —** using phylogenies as analytical frameworks for rigorous understanding of the evolution of various traits or conditions of interest

Common Phylogenetic Tree Terminology



Terminal Nodes

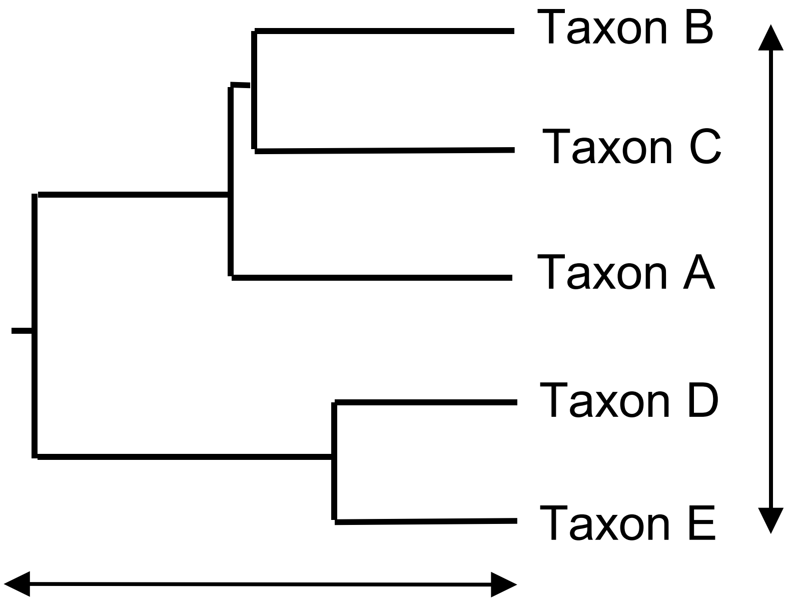
Branches or Lineages

Represent the TAXA (genes, populations, species, etc.) used to infer the phylogeny

Ancestral Node or ROOT of the Tree

Internal Nodes or Divergence Points (represent hypothetical ancestors of the taxa)

Phylogenetic trees diagram the *evolutionary relationships* between the taxa



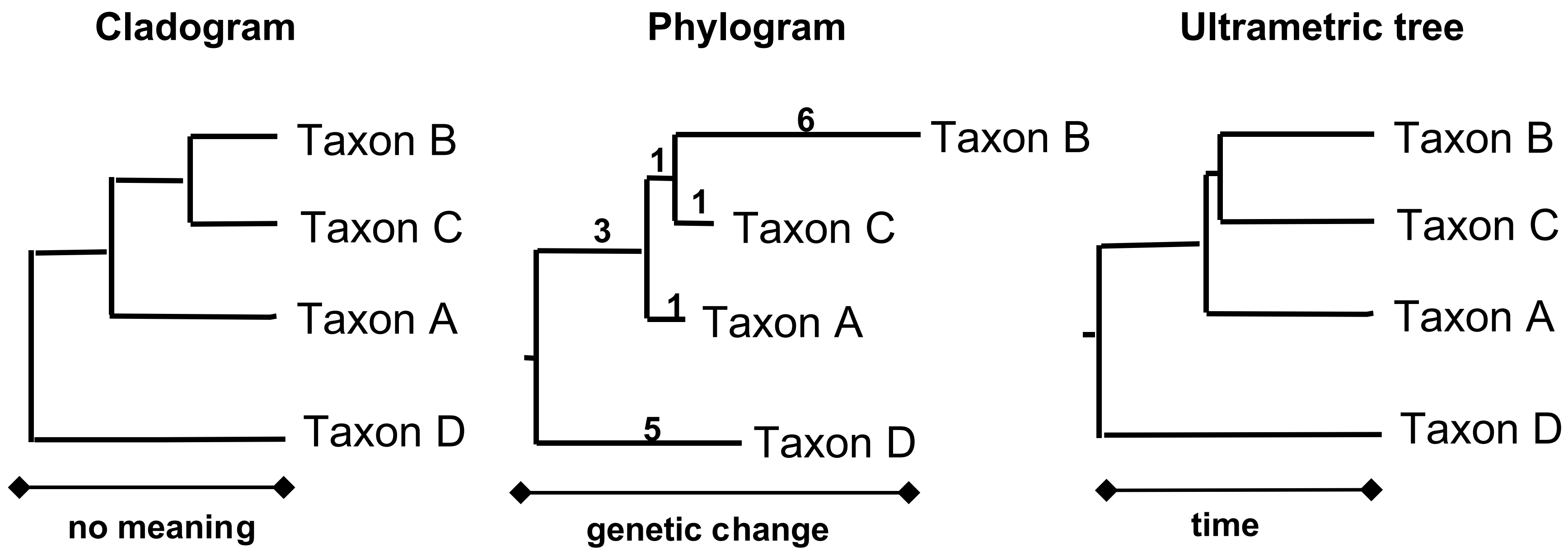
No meaning to the spacing between the taxa, or to the order in which they appear from top to bottom.

This dimension either can have no scale (for ‘cladograms’), can be proportional to genetic distance or amount of change (for ‘phylograms’ or ‘additive trees’), or can be proportional to time (for ‘ultrametric trees’ or true evolutionary trees).

((A,(B,C)),(D,E)) = The above phylogeny as nested parentheses

These say that B and C are more closely related to each other than either is to A, and that A, B, and C form a clade that is a sister group to the clade composed of D and E. If the tree has a time scale, then D and E are the most closely related.

Three types of trees

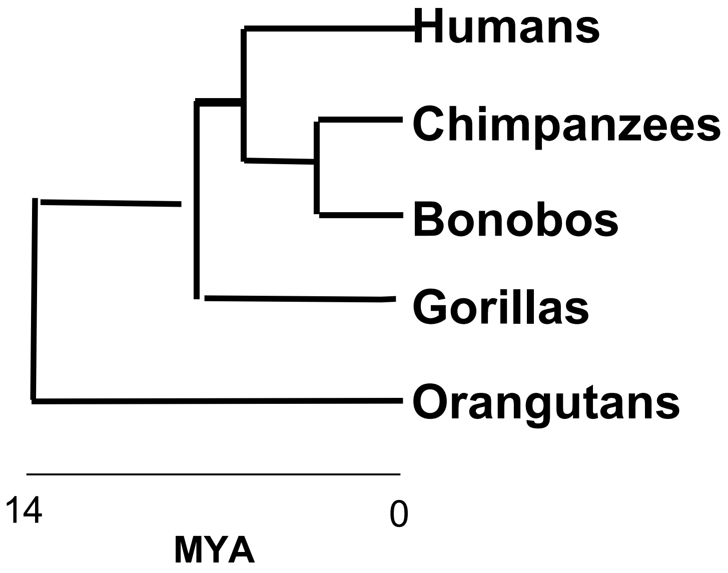


All show the same evolutionary relationships, or branching orders, between the taxa.

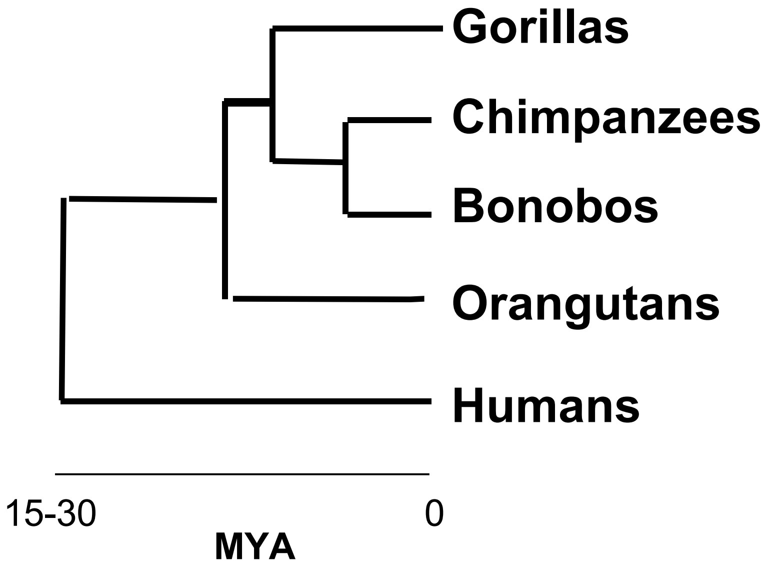
A few examples of what can be inferred from phylogenetic trees built from DNA or protein sequence data:

- Which species are the closest living relatives of modern humans?
- Did the infamous *Florida Dentist* infect his patients with HIV?
- What were the origins of specific transposable elements?
- Plus countless others…..

Which species are the closest living relatives of modern humans?



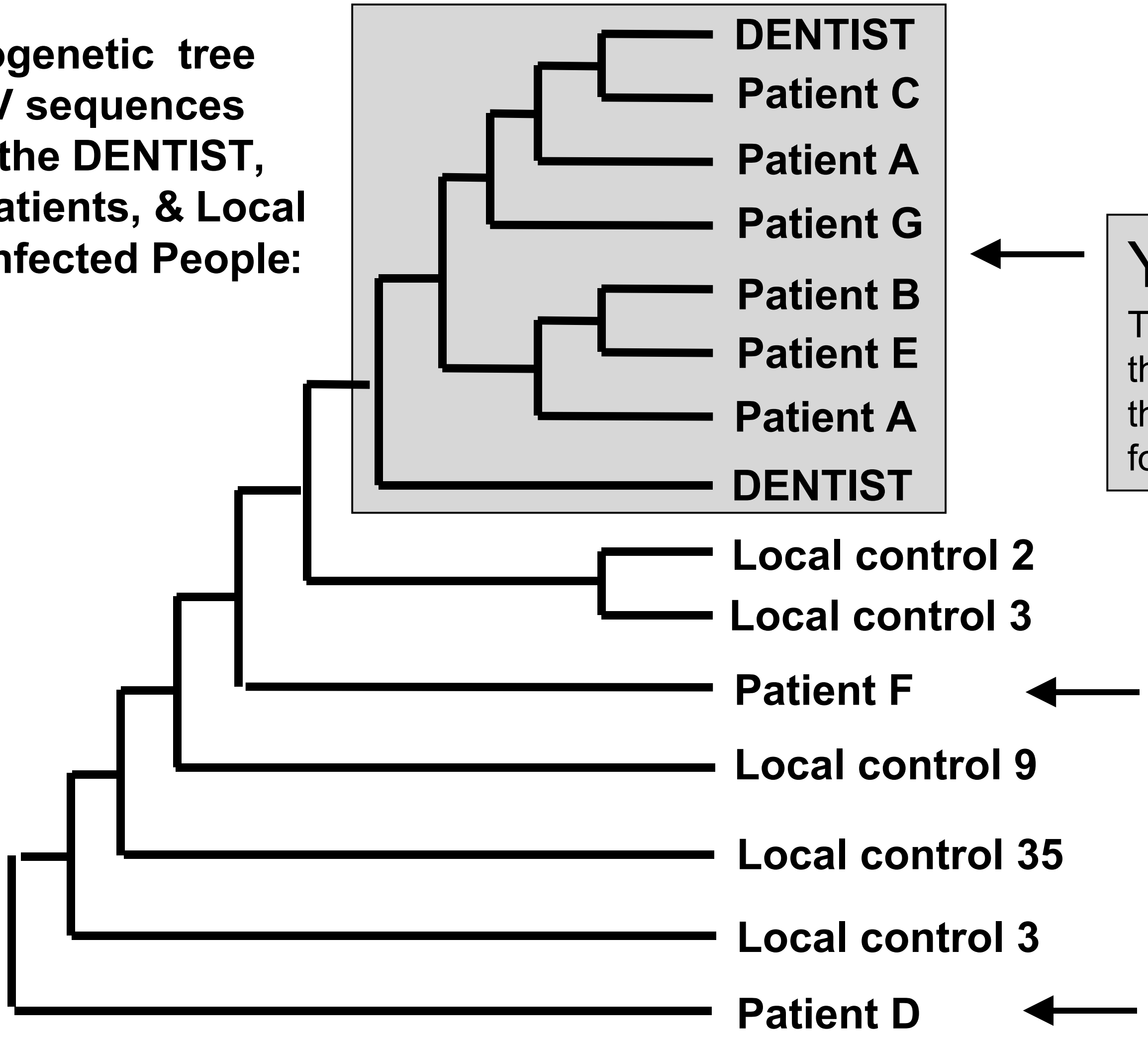
Mitochondrial DNA, most nuclear DNA-encoded genes, and DNA/DNA hybridization all show that bonobos and chimpanzees are related more closely to humans than either are to gorillas.



The pre-molecular view was that the great apes (chimpanzees, gorillas and orangutans) formed a clade separate from humans, and that humans diverged from the apes at least 15-30 MYA.

Did the *Florida Dentist* infect his patients with HIV?

Phylogenetic tree of HIV sequences from the DENTIST, his Patients, & Local HIV-infected People:



Yes:

The HIV sequences from these patients fall within the clade of HIV sequences found in the dentist.

No

No

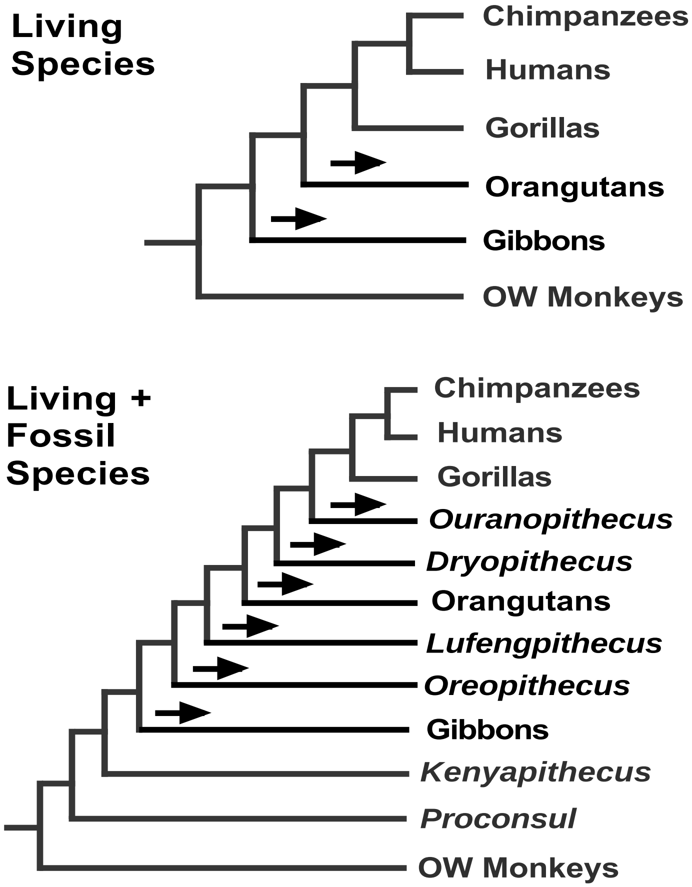
From Ou et *al*. (1992) and Page & Holmes (1998)

A few examples of what can be learned from character analysis using phylogenies as analytical frameworks:

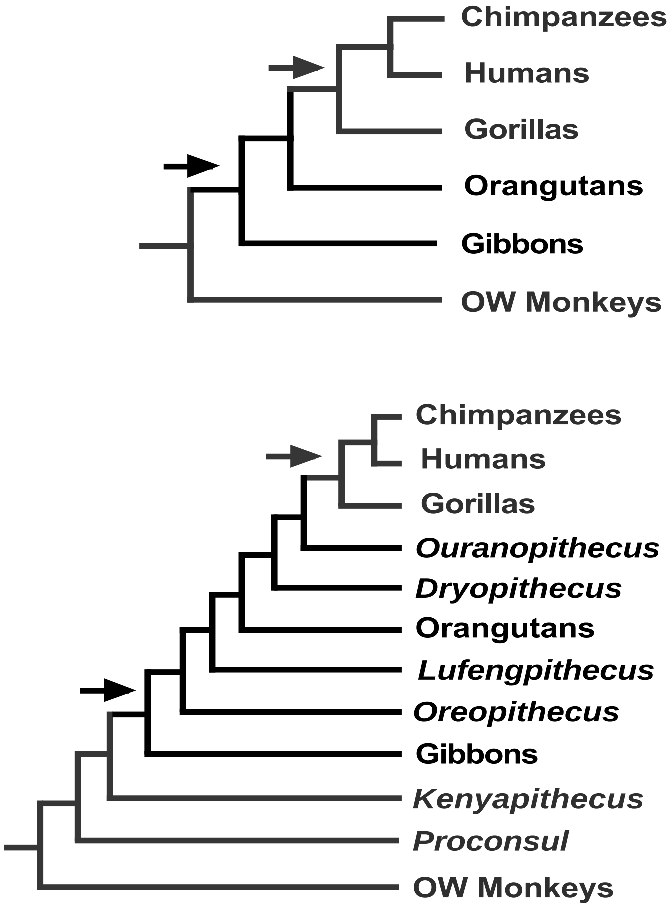
- When did specific episodes of positive Darwinian selection occur during evolutionary history?
- Which genetic *changes* are unique to the human lineage?
- What was the most likely geographical location of the common ancestor of the African apes and humans?
- Plus countless others…..

What was the most likely geographical location of the common ancestor of the African apes and humans?

Scenario A: Africa as species fountain



Scenario B: Eurasia as ancestral homeland

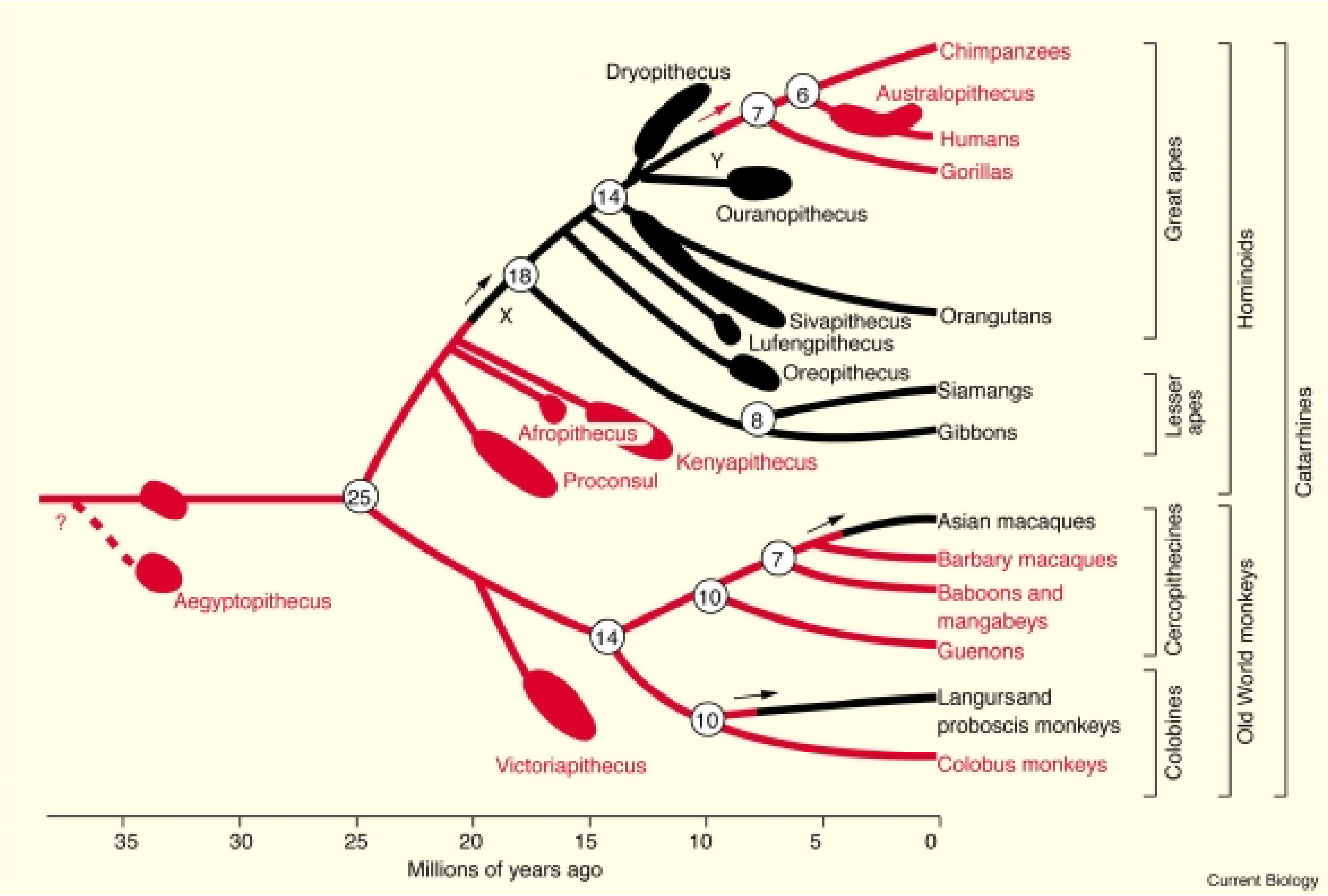


Modified from: Stewart, C.-B. & Disotell, T.R. (1998) *Current Biology 8*: R582-588.

Eurasia = Black
Africa = Red
→ = Dispersal

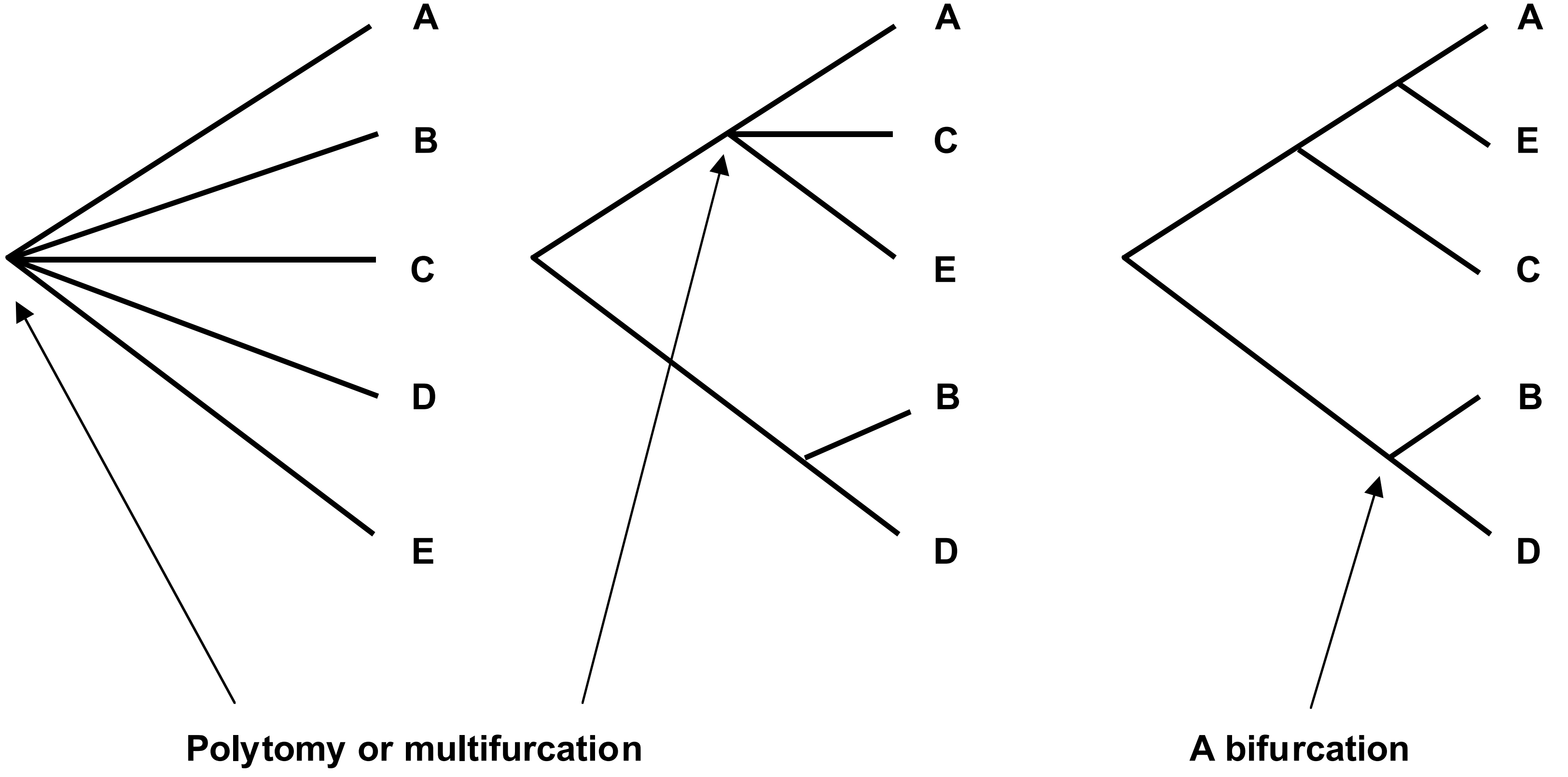
Scenario B requires four fewer dispersal events

Inferred ancestral dispersal patterns of primates between Africa and Eurasia



From: Stewart, C.-B. & Disotell, T.R. (1998) *Current Biology 8*: R582-588.

The goal of phylogeny inference is to resolve the branching orders of lineages in evolutionary trees:



Completely unresolved or "star" phylogeny

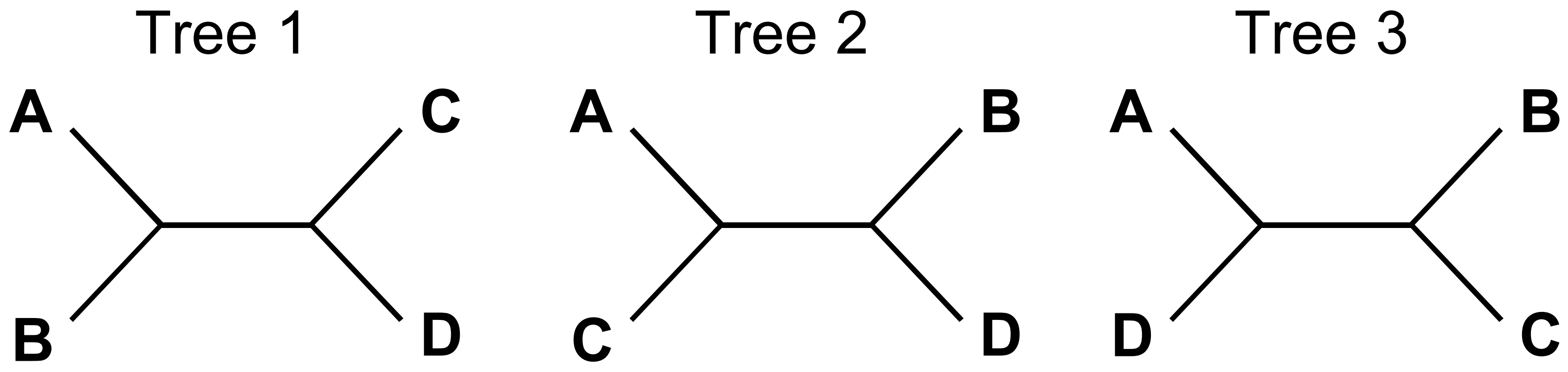
Partially resolved phylogeny

Fully resolved, bifurcating phylogeny

Polytomy or multifurcation

A bifurcation

There are three possible unrooted trees for four taxa (A, B, C, D)



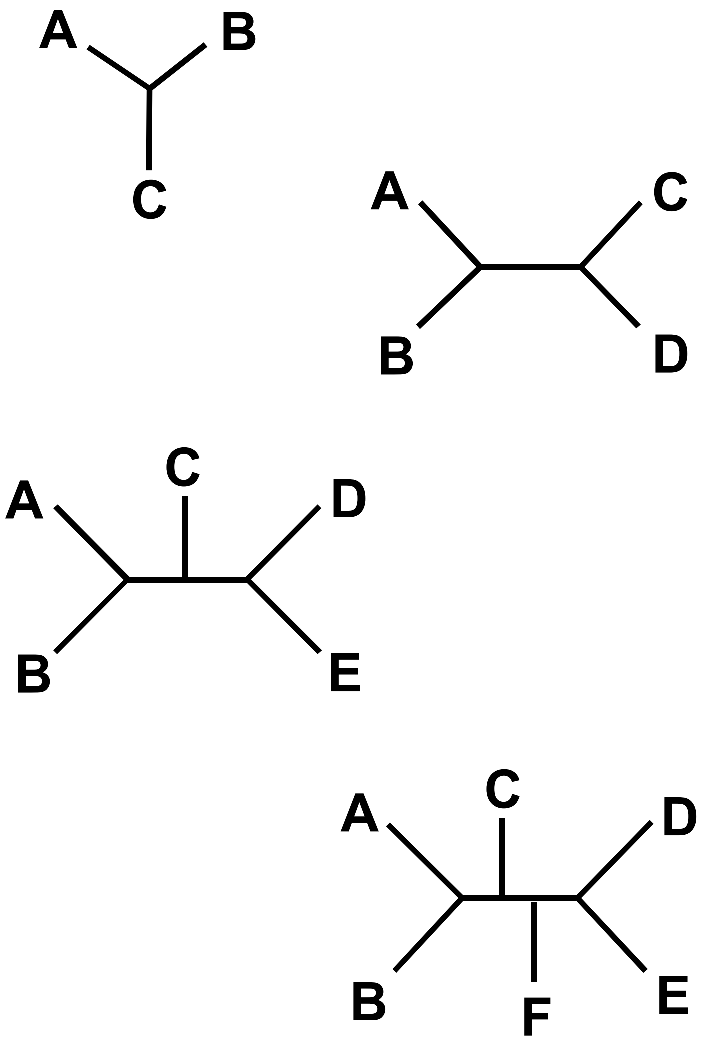
Tree 1

Tree 2

Tree 3

Phylogenetic tree building (or inference) methods are aimed at discovering which of the possible unrooted trees is "correct". We would like this to be the “true” biological tree — that is, one that accurately represents the evolutionary history of the taxa. However, we must settle for discovering the *computationally correct* or *optimal* tree for the phylogenetic method of choice.

The number of unrooted trees increases in a greater than exponential manner with number of taxa

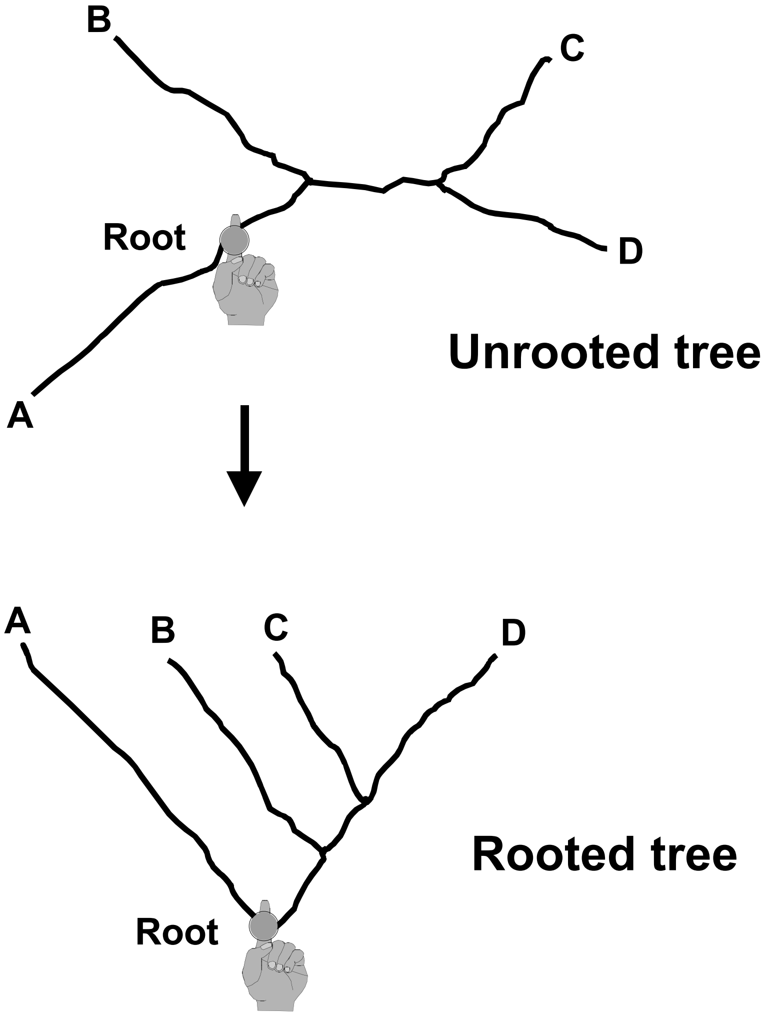


# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.
.	.
30	$\approx 3.58 \times 10^{36}$

(2N - 5)!! = # unrooted trees for N taxa

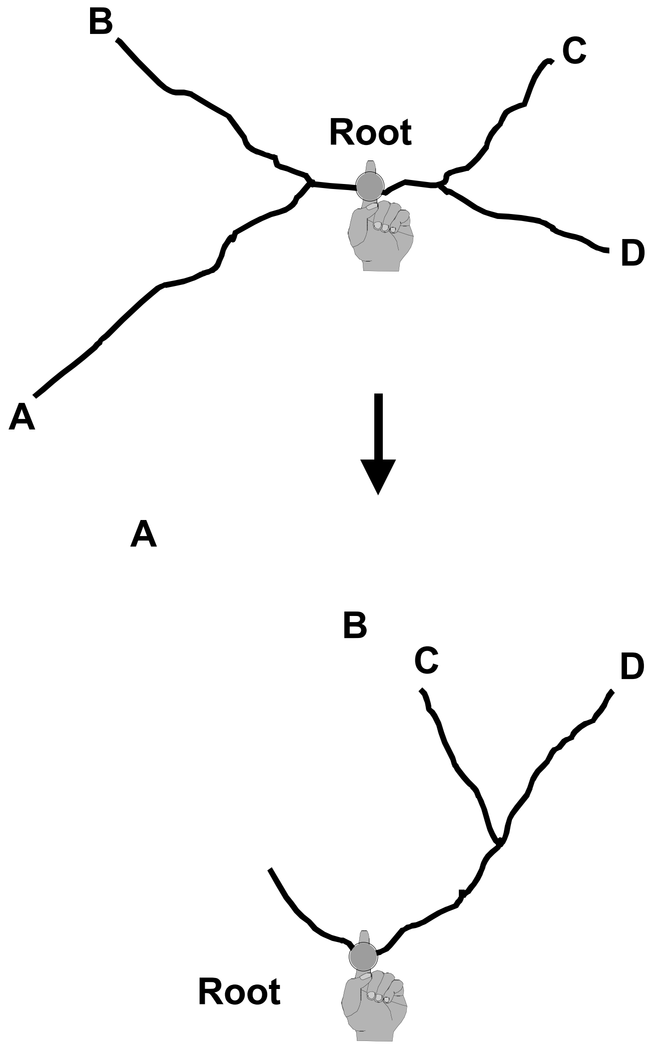
Inferring evolutionary *relationships* between the taxa requires rooting the tree:

To root a tree mentally, imagine that the tree is made of string. Grab the string at the root ● and tug on it until the ends of the string (the taxa) fall opposite the root:



Note that in this rooted tree, taxon A is no more closely related to taxon B than it is to C or D.

Now, try it again with the root at another position:



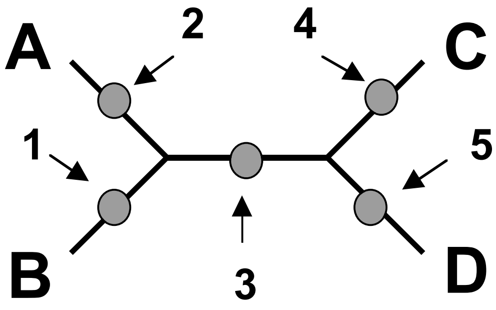
Unrooted tree

Rooted tree

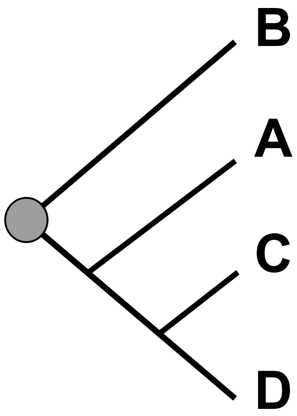
Note that in this rooted tree, taxon A is most closely related to taxon B, and together they are equally distantly related to taxa C and D.

An unrooted, four-taxon tree theoretically can be rooted in five different places to produce five different rooted trees

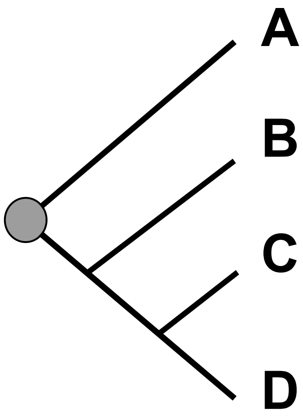
The unrooted tree 1:



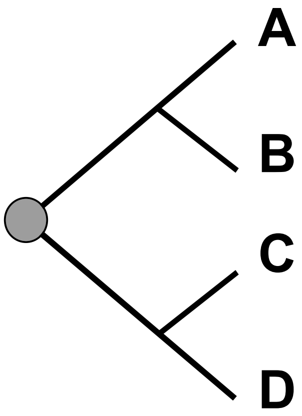
Rooted tree 1a



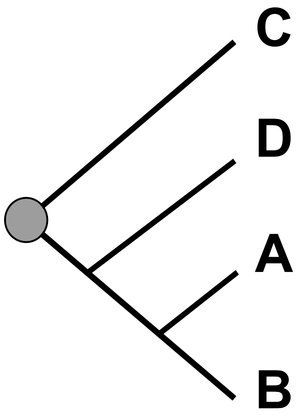
Rooted tree 1b



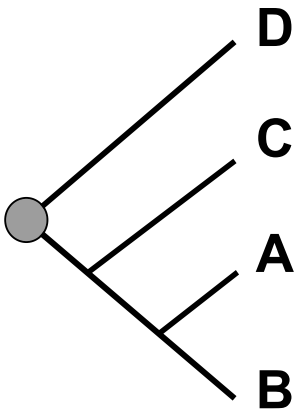
Rooted tree 1c



Rooted tree 1d

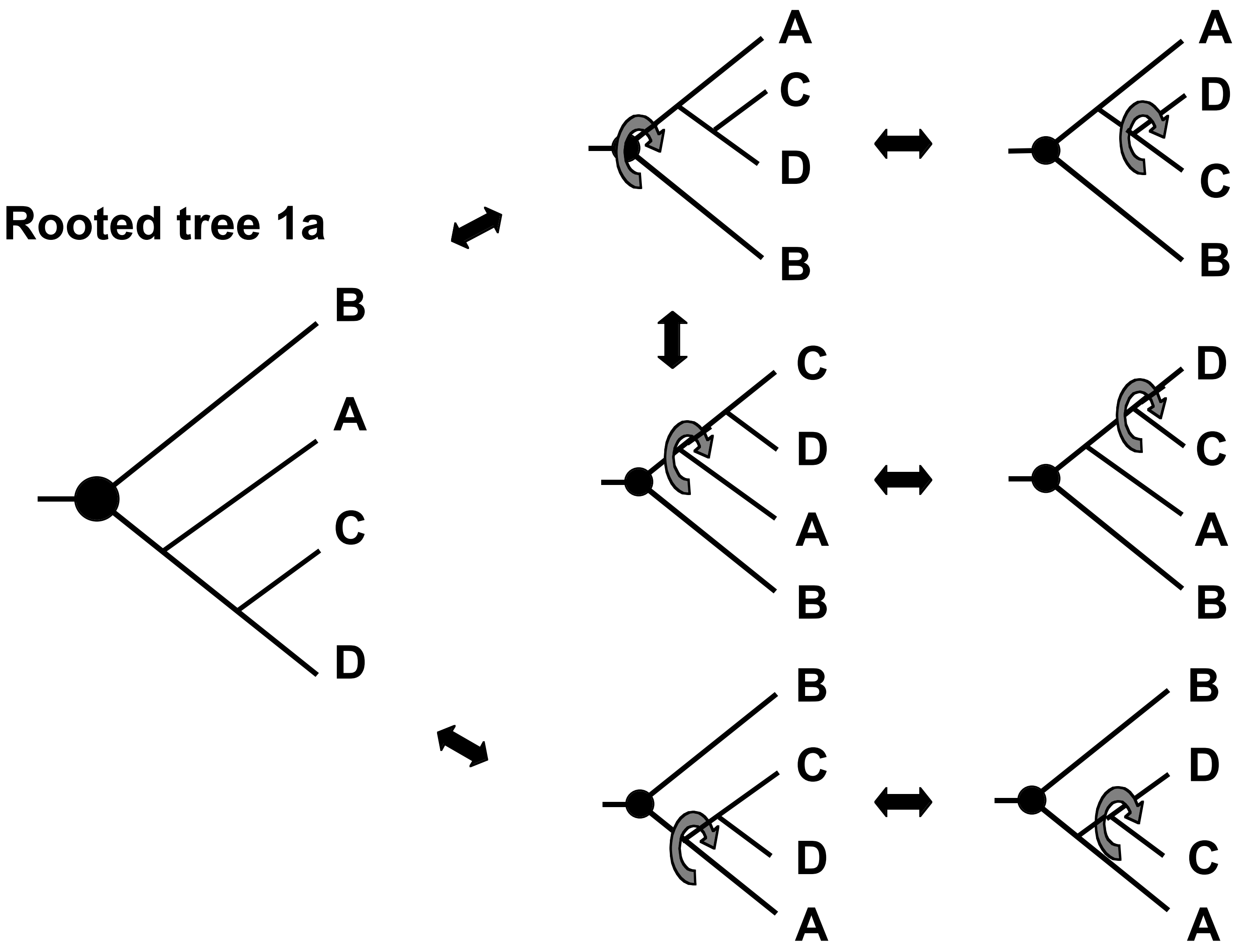


Rooted tree 1e



These trees show five different evolutionary relationships among the taxa!

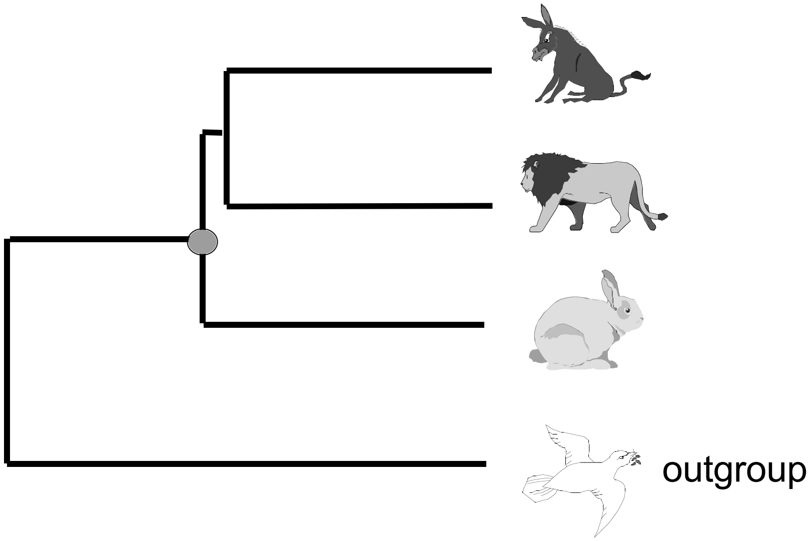
All of these rearrangements show the same evolutionary relationships between the taxa



There are two major ways to root trees:

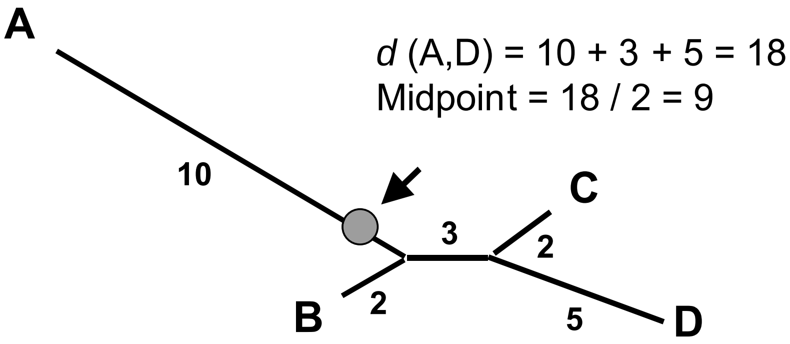
By outgroup:

Uses taxa (the “outgroup”) that are known to fall outside of the group of interest (the “ingroup”). Requires some prior knowledge about the relationships among the taxa. The outgroup can either be species (*e.g.*, birds to root a mammalian tree) or previous gene duplicates (*e.g.*, α-globins to root β-globins).



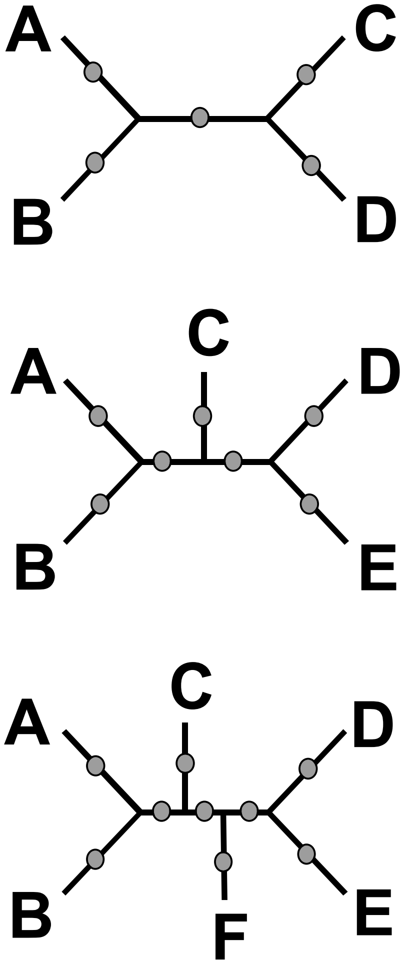
By midpoint or distance:

Roots the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. Assumes that the taxa are evolving in a clock-like manner. This assumption is built into some of the distance-based tree building methods.



d (A,D) = 10 + 3 + 5 = 18
Midpoint = 18 / 2 = 9

Each unrooted tree theoretically can be rooted anywhere along any of its branches



# Taxa	# Unrooted Trees	x	● # Roots	=	# Rooted Trees
3	1		3		3
4	3		5		15
5	15		7		105
6	105		9		945
7	945		11		10,395
8	10,935		13		135,135
9	135,135		15		2,027,025
.	.		.		.
.	.		.		.
.	.		.		.
.	.		.		.
30	$\sim3.58\text{x}10^{36}$		57		$\sim2.04\text{x}10^{38}$

(2N - 3)!! = # unrooted trees for N taxa

Molecular phylogenetic tree building methods:

Are mathematical and/or statistical methods for inferring the divergence order of taxa, as well as the lengths of the branches that connect them. There are many phylogenetic methods available today, each having strengths and weaknesses. Most can be classified as follows:

		COMPUTATIONAL METHOD	
		Optimality criterion	Clustering algorithm
DATA TYPE	Characters	PARSIMONY MAXIMUM LIKELIHOOD	
	Distances	MINIMUM EVOLUTION LEAST SQUARES	UPGMA NEIGHBOR-JOINING

Types of data used in phylogenetic inference:

Character-based methods: Use the aligned characters, such as DNA or protein sequences, directly during tree inference.

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

Distance-based methods: Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

Example 1: Uncorrected "p" distance (=observed percent sequence difference)

Example 2: Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)

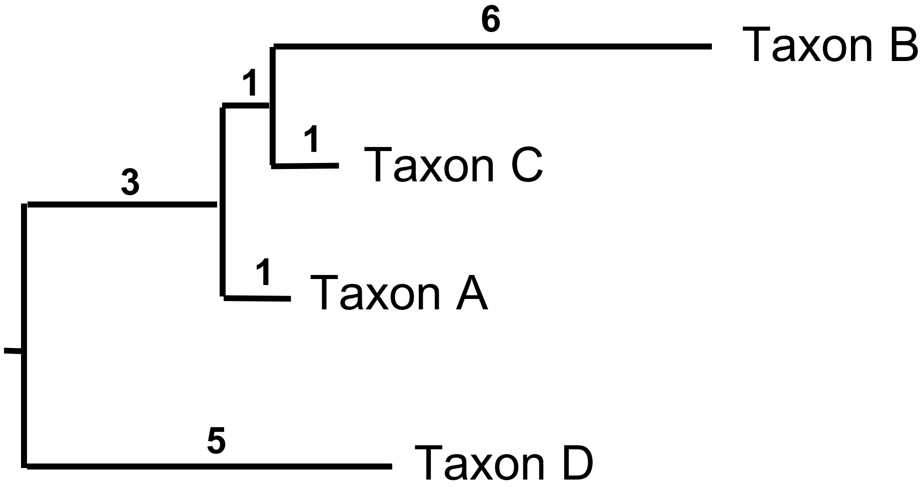
Similarity *vs.* Evolutionary Relationship:

Similarity and relationship are *not* the same thing, even though evolutionary relationship is inferred from certain types of similarity.

Similar: having likeness or resemblance (an observation)

Related: genetically connected (an historical fact)

Two taxa can be most similar without being most closely-related:



C is more similar in sequence to A (*d* = 3) than to B (*d* = 7), but C and B are most closely related (that is, C and B shared a common ancestor more recently than either did with A).

Types of Similarity

Observed similarity between two entities can be due to:

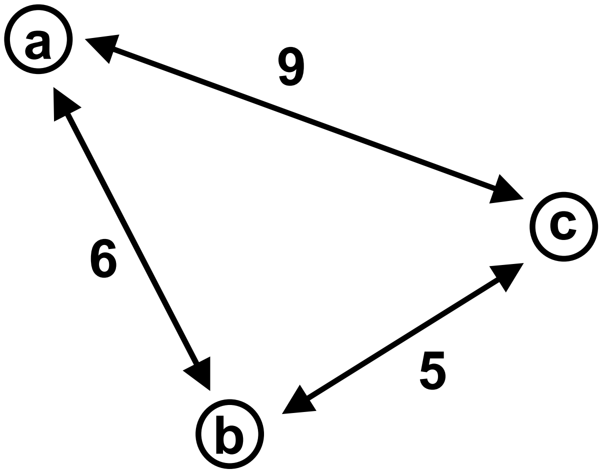
Evolutionary relationship:
Shared ancestral characters (‘plesiomorphies’)
Shared derived characters (‘’synapomorphy’)

Homoplasy (independent evolution of the same character):
Convergent events (in either related on unrelated entities)
Parallel events (in related entities)
Reversals (in related entities)

Character-based methods can tease apart types of similarity and theoretically find the true evolutionary tree. Similarity = relationship only if certain conditions are met (if the distances are ‘ultrametric’).

METRIC DISTANCES between any two or three taxa (a, b, and c) have the following properties:

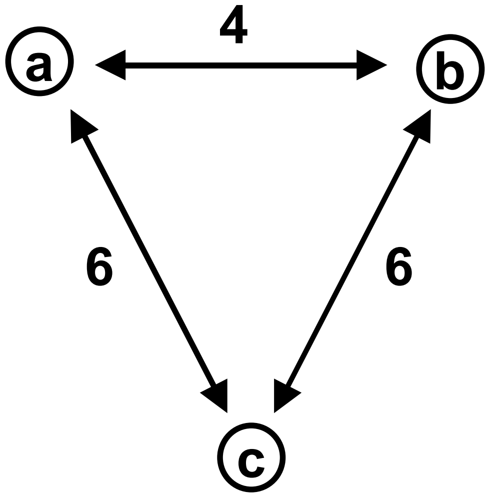
Property 1:	$d\ (a, b) \geq 0$	Non-negativity
Property 2:	$d\ (a, b) = d\ (b, a)$	Symmetry
Property 3:	$d\ (a, b) = 0$ if and only if $a = b$	Distinctness
and...		
Property 4:	$d\ (a, c) \leq d\ (a, b) + d\ (b, c)$	Triangle inequality:



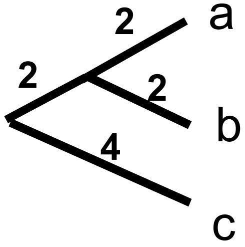
ULTRAMETRIC DISTANCES
must satisfy the previous four conditions, plus:

Property 5 ***d*** **(a, b) ≤ maximum [*d* (a, c), *d* (b, c)]**

This implies that the two largest distances are equal, so that they define an isosceles triangle:



Similarity = Relationship if the distances are ultrametric!



If distances are ultrametric, then the sequences are evolving in a perfectly clock-like manner, thus can be used in UPGMA trees and for the most precise calculations of divergence dates.

ADDITIVE DISTANCES:

Property 6:

$$d\ (\mathrm{a, b}) + d\ (\mathrm{c, d}) \leq \text{maximum}\ [d\ (\mathrm{a, c}) + d\ (\mathrm{b, d}), d\ (\mathrm{a, d}) + d\ (\mathrm{b, c})]$$

For distances to fit into an evolutionary tree, they must be either metric or ultrametric, and they must be additive. Estimated distances often fall short of these criteria, and thus can fail to produce correct evolutionary trees.

Types of computational methods:

Clustering algorithms: Use pairwise distances. Are purely algorithmic methods, in which the algorithm itself defines the the tree selection criterion. Tend to be very fast programs that produce singular trees rooted by distance. No objective function to compare to other trees, even if numerous other trees could explain the data equally well. Warning: Finding a singular tree is not necessarily the same as finding the "true” evolutionary tree.

Optimality approaches: Use either character or distance data. First define an *optimality criterion* (minimum branch lengths, fewest number of events, highest likelihood), and then use a specific algorithm for finding trees with the best value for the objective function. Can identify many equally optimal trees, if such exist. Warning: Finding an optimal tree is not necessarily the same as finding the "true” tree.

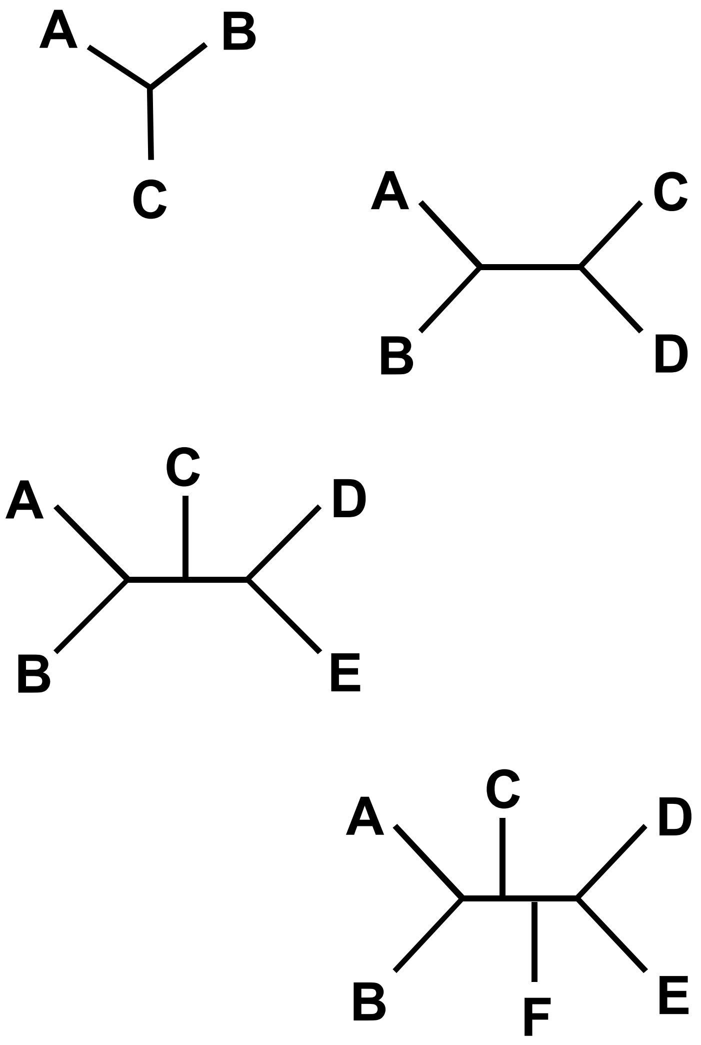
Computational methods for finding optimal trees:

Exact algorithms: "Guarantee" to find the optimal or "best" tree for the method of choice. Two types used in tree building:

> **Exhaustive search:** Evaluates all possible unrooted trees, choosing the one with the best score for the method.
>
> **Branch-and-bound search:** Eliminates the parts of the search tree that only contain suboptimal solutions.

Heuristic algorithms: Approximate or “quick-and-dirty” methods that attempt to find the optimal tree for the method of choice, but cannot guarantee to do so. Heuristic searches often operate by “hill-climbing” methods.

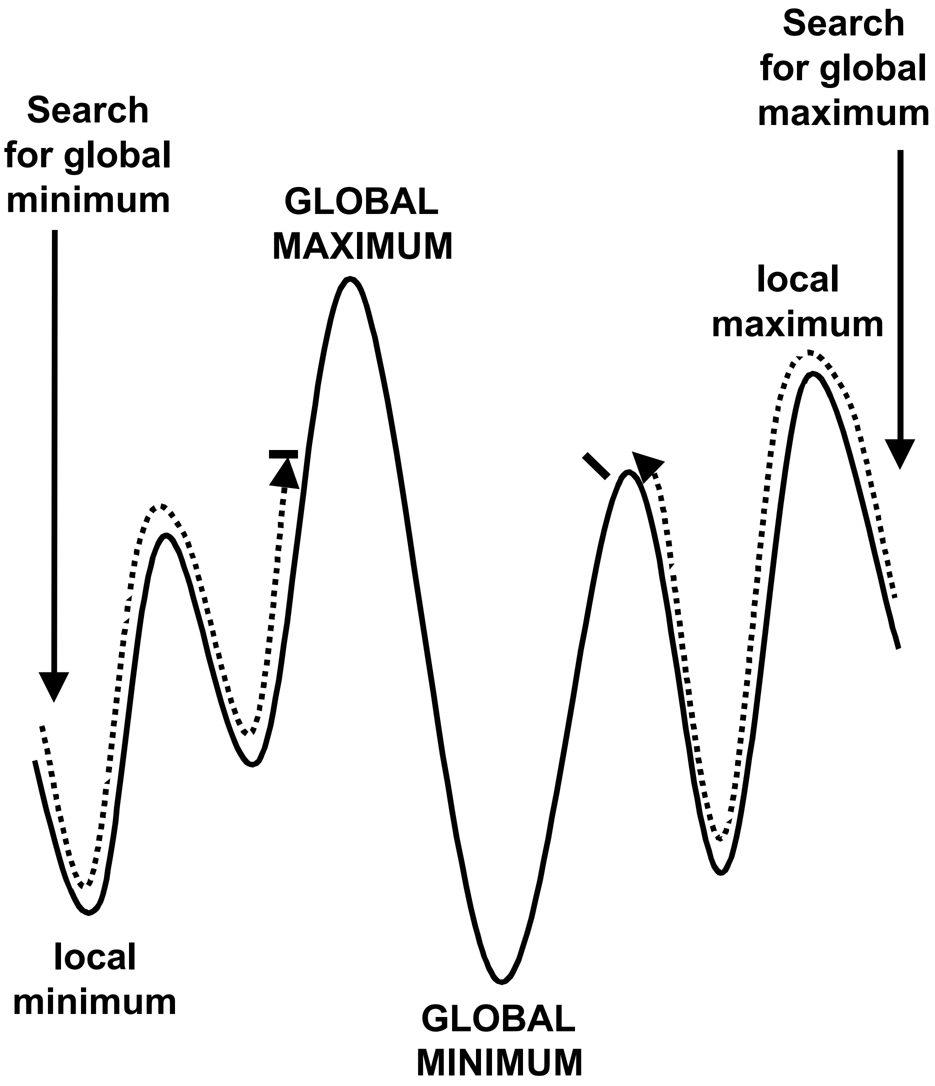
Exact searches become increasingly difficult, and eventually impossible, as the number of taxa increases:



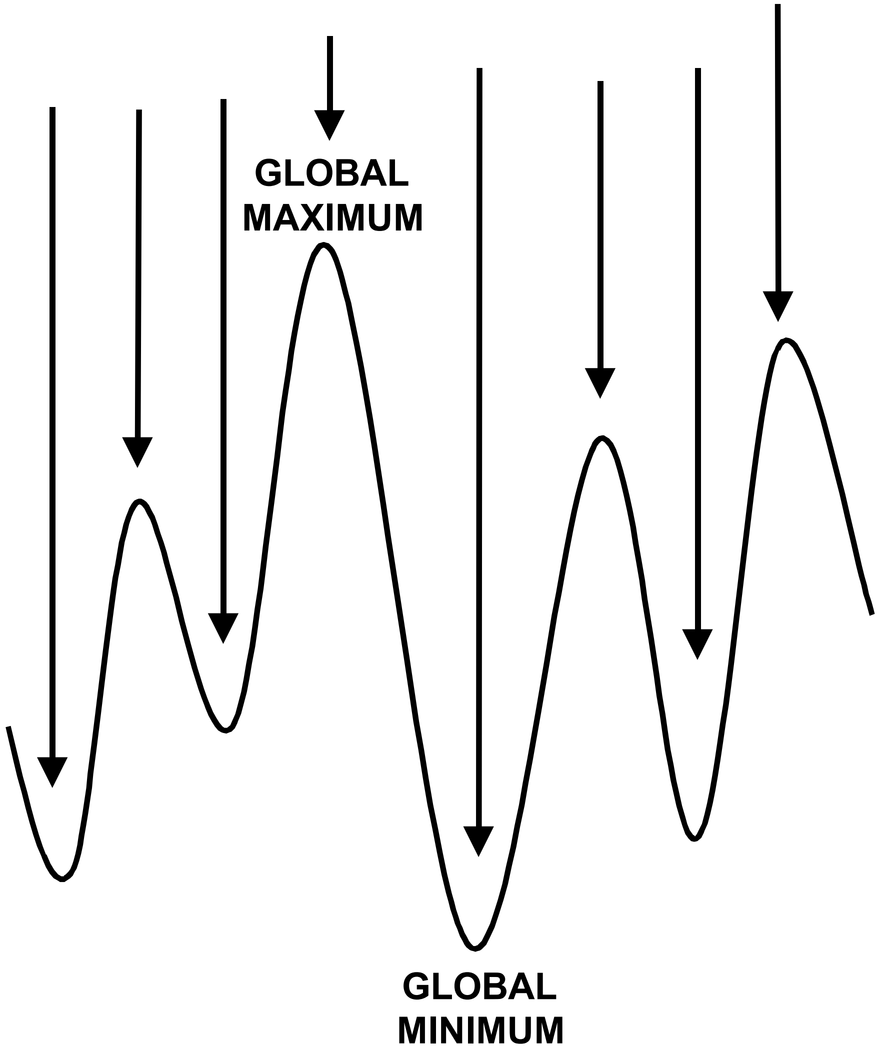
# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.
.	.
30	≈3.58 x 10^{36}

(2N - 5)!! = # unrooted trees for N taxa

Heuristic search algorithms are input order dependent and can get stuck in local minima or maxima



Rerunning heuristic searches using different input orders of taxa can help find global minima or maxima



Classification of phylogenetic inference methods

		COMPUTATIONAL METHOD	
		Optimality criterion	Clustering algorithm
DATA TYPE	Characters	PARSIMONY MAXIMUM LIKELIHOOD	
	Distances	MINIMUM EVOLUTION LEAST SQUARES	UPGMA NEIGHBOR-JOINING

Parsimony methods:

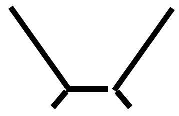
Optimality criterion: The 'most-parsimonious' tree is the one that requires the fewest number of evolutionary events (*e.g.,* nucleotide substitutions, amino acid replacements) to explain the sequences.

Advantages:

- Are simple, intuitive, and logical (many possible by 'pencil-and-paper').
- Can be used on molecular *and* non-molecular (*e.g.,* morphological) data.
- Can tease apart types of similarity (shared-derived, shared-ancestral, homoplasy)
- Can be used for character (can infer the exact substitutions) and rate analysis.
- Can be used to infer the sequences of the extinct (hypothetical) ancestors.

Disadvantages:

- Are simple, intuitive, and logical (derived from "Medieval logic", not statistics!)
- Can be fooled by high levels of homoplasy ('same' events).
- Can become positively misleading in the "Felsenstein Zone":



[See Stewart (1993) for a simple explanation of parsimony analysis, and Swofford *et al.* (1996) for a detailed explanation of various parsimony methods.]

Maximum likelihood (ML) methods

Optimality criterion: ML methods evaluate phylogenetic hypotheses in terms of the probability that a proposed model of the evolutionary process and the proposed unrooted tree would give rise to the observed data. The tree found to have the highest ML value is considered to be the preferred tree.

Advantages:

- Are inherently statistical and evolutionary model-based.
- Usually the most ‘consistent’ of the methods available.
- Can be used for character (can infer the exact substitutions) and rate analysis.
- Can be used to infer the sequences of the extinct (hypothetical) ancestors.
- Can help account for branch-length effects in unbalanced trees.
- Can be applied to nucleotide or amino acid sequences, and other types of data.

Disadvantages:

- Are not as simple and intuitive as many other methods.
- Are computationally very intense (limits number of taxa and length of sequence).
- Like parsimony, can be fooled by high levels of homoplasy.
- Violations of the assumed model can lead to incorrect trees.

Minimum evolution (ME) methods

Optimality criterion: The tree(s) with the shortest sum of the branch lengths (or overall tree length) is chosen as the best tree.

Advantages:

- Can be used on indirectly-measured distances (immunological, hybridization).
- Distances can be ‘corrected’ for unseen events.
- Usually faster than character-based methods.
- Can be used for some rate analyses.
- Has an objective function (as compared to clustering methods).

Disadvantages:

- Information lost when characters transformed to distances.
- *Cannot* be used for character analysis.
- Slower than clustering methods.

Clustering methods (UPGMA & N-J)

Optimality criterion: NONE. The algorithm itself builds 'the' tree.

Advantages:

- Can be used on indirectly-measured distances (immunological, hybridization).
- Distances can be 'corrected' for unseen events/
- The fastest of the methods available (N-J is screamingly fast!).
- Can therefore analyze very large datasets quickly (needed for HIV, etc.).
- Can be used for some types of rate and date analysis.

Disadvantages:

- Similarity and relationship are not necessarily the same thing, so clustering by similarity does not necessarily give an evolutionary tree.
- *Cannot* be used for character analysis!
- Have no explicit optimization criteria, so one cannot even know if the program worked properly to find the correct tree for the method.

Recommended Readings in Phylogenetic Inference (or "Tree Building")

Roderick D.M. Page & Edward C. Holmes (1998) *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science Ltd., Oxford.

This a GREAT 'primer' on molecular evolution! Chapters 2, 5 & 6 are highly recommended for explaining phylogenetic trees.

Swofford, DL, Olsen, GJ, Waddell, PJ & Hillis, DM (1996) "Phylogenetic Inference", pp. 407-514 in Molecular Systematics, DM Hillis, C Moritz & BK Mable, eds. Sinauer Associates, Sunderland MA.

Hillis, DM, Mable, BK & Moritz, C (1996) "Applications of Molecular Systematics: The State of the Field and a Look to the Future", pp. 515-543 in *Molecular Systematics*, DM Hillis, C Moritz & BK Mable, eds. Sinauer Associates, Sunderland MA.

These are more advanced reviews about phylogenetic methods, and are highly recommended for serious practitioners.

Recommended Readings in Character and Rate Analysis

Roderick D.M. Page & Edward C. Holmes (1998) ***Molecular Evolution: A Phylogenetic Approach*****. Blackwell Science Ltd., Oxford.**

Chapters 7 & 8 are recommended for these purposes.

Maddison, D.R & Maddison, W.P. (2000) ***MacClade 4: Analysis of Phylogeny and Character Evolution.*** **Sinauer Associates, Sunderland, MA.**

The user’s manual has much valuable background and information about character analysis.

Highly Recommended Programs for Phylogenetic Inference and Evolutionary Analysis

Swofford, D.M. (1998) *PAUP* 4: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

> This is the most versatile and user-friendly phylogenetic analysis package currently available. *PAUP** has parsimony, likelihood, and distance methods. It is sold for a nominal cost. Available for several platforms; the PowerMac version is fast and menu-driven.

Maddison, D.R & Maddison, W.P. (2000) *MacClade 4: Analysis of Phylogeny and Character Evolution.* Sinauer Associates, Sunderland, MA.

> This is a versatile and user-friendly program that aids greatly in character analysis of molecular (and other) data. One can readily 'build' trees by click-and-drop methods, and save them for further analyses. Available for Macintosh and MacOS emulators. Fun!

Yang, Z. (1998) *PAML: Phylogenetic Analysis using Maximum Likelihood.* [Available from the author or online.]

> This is the scientifically best program available for testing alternative models of molecular evolution in a phylogenetic ML framework. Is user-hostile, but worth the effort. Available for several platforms.

END

or Demonstrations?